Genomics 94 (2009) 269–277

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Method

An integrated approach to infer causal associations among gene expression, genotype variation, and disease

Eunjee Lee [a], Seoae Cho [a], Kyunga Kim [b], Taesung Park [a,b,*]

[a] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea
[b] Department of Statistics, Seoul National University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 13 November 2008
Accepted 14 June 2009
Available online 18 June 2009

Keywords:
Causal association
Gene expression
Genotype variation
Integration

ABSTRACT

Gene expression data and genotype variation data are now capable of providing genome-wide patterns across many different clinical conditions. However, the separate analysis of these data has limitations in elucidating the complex network of gene interactions underlying complex traits, such as common human diseases. More information about the identity of key driver genes of common diseases comes from integrating these two heterogeneous types of data. We developed a two-step procedure to characterize complex diseases by integrating genotype variation data and gene expression data. The first step elucidates the causal relationship among genetic variation, gene expression level, and disease. Based on the causal relationship determined at the first step, the second step identifies significant gene expression traits whose effects on disease status or whose responses to disease status are modified by the specific genotype variation. For the selected significant genes, a pathway enrichment analysis can be performed to identify the genetic mechanism of a complex disease. The proposed two-step procedure was shown to be an effective method for integrating three different levels of data, i.e., genotype variation, gene expression and disease status. By applying the proposed procedure to a chronic fatigue syndrome (CFS) dataset, we identified a list of potential causal genes for CFS, and found an evidence for difference in genetic mechanisms of the etiology between CFS without 'a major depressive disorder with melancholic features' (CFS) and CFS with 'a major depressive disorder with melancholic features' (CFS-MDD/m). Especially, the SNPs within NR3C1 gene were shown to differently influence the susceptibility of developing CFS and CFS-MDD/m through integrative action with gene expression levels.

Introduction

In the past few years, high through-put technologies, such as gene expression microarrays and genotyping techniques, have provided efficient ways to measure gene expression levels and genotype variation on genome-wide scale. Various approaches have been proposed to analyze gene expression data and genotype variation data, and hence to discover a complex network of biochemical process underlying living systems [1] and common human diseases [2,3], and candidate genes [4]. In the analysis of gene expression data, for example, the identification of differentially expressed genes between two groups has been of great interest, and has been conducted via various statistical tests [5–7]. In analyzing genotype variation data, logistic regression has been commonly used to model the relationship between binary clinical outcomes and discrete predictors, such as genotypes [8].

Despite the availability of different levels of genome-wide data, most studies have been based on a separate analysis of single-level data to unravel complex biological mechanisms, such as complex human diseases. Complex diseases can be explained at different levels of biological mechanisms, including DNA, gene expression and phenotype levels. While there is a separate mechanism at each level, the mechanisms at different levels are closely related to each other in initiating and influencing complex diseases. Furthermore, most common human diseases have complex etiology involving action of many genes as well as dynamic gene–environment interactions. Therefore, separate analyses of single-level data have a limitation in identifying and characterizing genes that are associated with susceptibility of complex human diseases. Integration of different types of data, such as gene expression, genotype variation and clinical outcomes, can provide more comprehensive information on a complex disease, and hence can elucidate complex networks of gene interactions underlying complex traits.

Gene expression data have recently been combined with other experimental data, such as genotype variation data and clinical data, to facilitate identification of key mechanistic drivers of complex traits in mice [9,10] and human [11,12]. One way to combine gene expression and genotype variation data is to treat the expression level of a particular gene as a quantitative trait (so-called expression trait) in segregating populations, and to map chromosomal regions that

* Corresponding author.
E-mail addresses: eunjee01@gmail.com (E. Lee), sacho71@snu.ac.kr (S. Cho), kyunga.j.kim@gmail.com (K. Kim), tspark@stats.snu.ac.kr (T. Park).

0888-7543/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2009.06.002

control the expression trait. Those genetic regions are called expression quantitative trait loci (eQTL). This combined technique could reveal a remarkable wealth of quantitative heritable variation in the transcriptome, as shown in human and mouse [9,11,12]. However, the association between gene expression and a disease might be confounded by the genotype variation, especially when the risk of disease varies between genotype variations, and expression traits are also regulated by genotype variations. Due to the number of tests for eQTL approach, this approach leads to an increased number of false discoveries. There is one study [13] that demonstrated this problem and proposed a mixture over markers model that shares information across both markers and transcripts.

Other integration approaches for combining different types of data have been proposed [10,13–17]. Firstly, Schadt et al. [10] introduced an integration methodology to discover key driver of a complex trait by identifying correlated trait–gene pairs under the influence of common genetic loci, then using the genetics to help identify the direction of causality. The utility of this approach was demonstrated by experimentally validating that three newly identified causal genes are involved with the obesity susceptibility. The approach proposed by Schadt et al. includes multiple steps. First, pair-wise relationships among genotype variation, gene expression, and a complex trait, are respectively investigated by identifying quantitative trait loci (QTL) for the complex trait, selecting gene expression traits correlated with the complex trait, and detecting expression QTL (eQTL), which overlap the identified QTL, for the selected expression traits. In the final step, the likelihood based causality model selection (LCMS) test is used to identify the causal relationships of the genes detected with overlapping loci.

Their filtering step is effective in reducing the large search-space problem, especially in eQTL search and LCMS testing. However, it might result in more false negatives than exhaustive searches like our method, in detecting causal relationships of the genes, especially when a true causal relationship exists based on the interaction effects among genotype, gene expression and a trait of interest, but any pairwise association is weak. It would be partially because there is no guarantee that the overlapping loci are directly related to those interaction effects, and hence the approach of Schadt et al. does not comprehensively handle the interaction effects. Also, they focused on the causal associations to determine which of the genes causes disease, rather than responds to the disease state, among many genes whose expression changes are associated with disease traits.

Secondly, Lan et al. [17] proposed an approach to expose key components of gene regulatory networks by combining correlation analysis with linkage mapping. They annotated gene function for genes whose expressions are highly correlated with transcripts, which has significant expression quantitative trait locus/loci (eQTL). However, they did not discriminate causality relationships among expression level, genotype and disease. Thirdly, Kendziorski et al. [13] proposed empirical Bayes Mixture over Markers (MOM) model that shares information across both markers and transcripts. This model was proved useful in improving the specificity of eQTL identifications. This work used only genotype variation and gene expression data rather than disease status or trait data.

More recently, two studies proposed the network based or module based work, which allowed them to construct a more comprehensive view of the gene interactions underlying complex traits of interest [14,16]. Based on Weighted Gene Co-expression Network Analysis (WGCNA)[15], Presson et al. [14] proposed an approach that constructs gene sets (modules) from the observed gene expression data, and then relate them to gene ontology information to study their biological plausibility. Their work includes steps to identify trait-module association and trait-related genetic marker association, but it does not provide the model-based statistical tests that our approach provides to examine the interaction effects among different types of data.

Note that this approach can alleviate the multiple testing problems. Another recent work by Chen et al. [16] identifies gene networks that are perturbed by susceptibility loci and that may in turn lead to disease. After constructing co-expression network combining gene expression data and genotype data, they identify sub-networks that are significantly associated with a complex of linked genetic loci related to disease traits. However, as the above approaches, these two studies did not consider the interaction effect between two different data types but our approach provides to examine the interaction effects among different types of data.

In this study, we developed a two-step procedure to conduct an integrative analysis of gene expression data, genetic variation data, and clinical data, which employs the LCMS test. Despite a similarity in using the same causality models, our approach differs from that of Schadt et al. in many aspects. We propose to select the most appropriate causality model for every combination of all SNPs and all genes profiled at the first step, based on the LCMS test. In the following second step, we identify gene expression traits whose effects on disease (i.e., candidate causal genes) or responses to disease (i.e., candidate reactive genes) are modified by the genotype variation, by employing more sophisticated statistical models that correspond to the causal relationships found in the first step. Our integrative approach has two major advantages. First, the proposed method is based on an exhaustive search while the Schadt et al.'s method is based on a non-exhaustive search. In our approach, the exhaustive investigation on causal relationships of all possible SNP–gene combinations can avoid false negatives which would result from a non-exhaustive search. Our approach can hence be more beneficial when employed in an exploratory (e.g., genome-wide) search, in which scientists want to identify more hypotheses to be further tested for biological significance. Second, our integrative approach allows one to examine the interaction effects among different types of data via the model-based statistical tests and hence to elucidate different disease susceptibility by identifying both candidate causal and reactive genes in a more effective and detailed manner. In other words, the model-based statistical tests enable one to infer both main and interaction effects via parameter estimation.

Results

The proposed two-step integrative analysis was applied to two CFS datasets, CFS with NF groups and CFS-MDD/m with NF groups, for investigating different etiological mechanisms of CFS. We also applied the multi-step procedure proposed by Schadt et al. [10] to the same datasets for the purpose of comparison.

Multi-step procedure by Schadt et al.

First, we carried out a gene expression analysis to detect differentially expressed genes across clinical outcomes We found that only a few genes are identified as differentially expressed (Table 1A) by three commonly used approaches such as the *t*-test, significance analysis of microarray (SAM) [5] and the Bayesian regression approach [6]. Second, genotype variation data and clinical outcomes were analyzed via logistic regression to detect the susceptibility genes of disease. Out of all 41 markers tested (Supplementary Table 1), nine markers were detected with significant genotype effect on initiation of CFS at a 5% significance level, while only four markers were detected with 5% false discovery rate (FDR [18]) (Table 1B). Interestingly, different sets of susceptible genes were identified as having statistically significant association with CFS and CFS-MDD/m. From the CFS vs. NF comparison, the seven markers in the NR3C1 gene were identified as significant markers linked to CFS. On the other hand, the CFS-MDD/m vs. NF comparison revealed the two significant markers in the COMT gene.

Table 1
Parallel analyses for respective association of gene expression and genotype variation with disease status.

Dataset	*t*-test	SAM test	Bayesian model		
A. Number of genes with significant change in expression levels over different disease status, which were detected via *t*-test, SAM test and Bayesian model					
CFS vs. NF	1	2	0		
CFS-MDD/m vs. NF	1	1	0		
Gene	SNP[a]	Chromosome	Position (Mb)[b]	CFS vs. NF[c]	CFS-MDD/m vs. NF[d]
B. Significant genotype variation linked to disease loci, which were detected via logistic regression					
NR3C1[e]	rs2918419	5	142.641	**0.0104**	0.3950
	rs1866388	5	142.702	**0.0010**[f]	**0.0472**
	rs860458	5	142.739	**0.0104**	0.3950
	rs852977	5	146.642	**0.0035**[f]	0.1878
	rs6196	5	146.660	**0.0208**	0.6423
	rs6188	5	146.667	**0.0027**[f]	**0.0396**
	rs258750	5	146.674	**0.0035**[f]	0.1009
COMT[g]	rs933271	22	18.311	0.0649	**0.0025**
	rs5993882	22	18.317	0.4306	**0.0114**

As multiple filtering steps in Schadt et al.'s procedure, the separate analyses were conducted respectively on two datasets, CFS vs. NF groups and CFS-MDD/m vs. NF groups. Bold numbers indicate p-values < 0.05.

[a] NCBI dbSNP Build number is 125 using Human Genome Build 35.1.
[b] Position of SNP on chromosome.
[c] p-value from logistic regression with CFS vs. NF data.
[d] p-value from logistic regression with CFS-MDD/m vs. NF data.
[e] Glucocorticoid receptor located at 5q34.
[f] Significant at the 5% false discovery rate (FDR) [36].
[g] Catechol-O-methyltransferase located at 22q11.1.

Finally, for each of the differentially expressed genes across clinical outcomes, eQTL were searched at each of the markers that were identified at the second step, via one-way ANOVA of genotype variation and gene expression data. We found no significant association between gene expression level and genotype variation for any genotype–gene expression combination at a 5% significant level. In other words, no significant results were detected for both datasets from the Schadt et al.'s multi-step method.

Two-step integrative analysis

For each combination of 20,160 genes and 41 SNPs, the proposed integrative analysis was conducted respectively on two datasets, CFS vs. NF groups and CFS-MDD/m vs. NF groups. For each gene–SNP combination, the best causal relationship was detected via the causality model selection at Step 1. In comparing CFS with NF groups, the reactive model was selected for ~70% of 20,160 genes on average, for all nine markers within two known CFS-related genes, such as NR3C1 and COMT (Table 2). However, in comparing CFS-MDD/m with NF groups, the causal model was selected for nearly 70% genes for three markers in the NR3C1 gene. This different tendency in the model selection results between CFS and CFS-MDD/m would imply different genetic mechanisms of CFS and CFS-MDD/m.

At Step 2, each gene–SNP combination data was analyzed based on one of the three statistical models, corresponding to the detected causal relationship. For all seven SNPs within NR3C1, significant causal relationships with gene expression levels were detected for either or both datasets (Table 2). Three SNPs (rs258750, rs6188 and rs852977) showed significant relationships with expression levels of a large number of genes, and can be candidates for genetic modulators of CFS-related regulatory pathways. We further conducted pathway enrichment analyses for these three SNPs, and present the results in the next section. In comparing CFS with NF groups, for the rs258750 marker, 105 genes were identified with differential expression across genotypes with 5% FDR from the independent test. This result is supported by the evidence of the neuroendocrine regulation of

Table 2
Two-step integration of genotype, gene expression and disease data based on causality model selection.

Gene	SNP	CFS vs. NF			CFS-MDD/m and NF		
		Causal[a]	Reactive[b]	Independent[c]	Causal[a]	Reactive[b]	Independent[c]
NR3C1[d]	rs2918419	0 (639)	7 (16,215)	0 (3306)	8 (13,955)	3 (5912)	0 (293)
	rs1866388	0 (165)	0 (16,872)	0 (3123)	15 (4136)	71 (15,976)	0 (48)
	rs860458	0 (639)	7 (16,215)	0 (3306)	8 (13,955)	3 (5912)	0 (293)
	rs852977	0 (230)	0 (17,001)	0 (2929)	**120 (9760)**	**73 (10,139)**	**0 (261)**
	rs6196	0 (604)	2 (15,037)	0 (4519)	0 (16,278)	0 (2013)	0 (1869)
	rs6188	0 (171)	7 (16,970)	1 (3019)	**52 (2939)**	**217 (17,074)**	**0 (147)**
	rs258750	**0 (242)**	**0 (16,279)**	**105 (3639)**	0 (2769)	14 (12,590)	0 (4801)
COMT[e]	rs933271	0 (1943)	0 (15,156)	0 (3061)	0 (169)	0 (16,872)	0 (3119)
	rs5993882	0 (1022)	0 (14,380)	0 (4758)	0 (547)	0 (17,333)	0 (2280)

The integrative analyses were conducted respectively on two datasets, CFS vs. NF groups and CFS-MDD/m vs. NF groups. Note that the results are presented only for nine SNPs within two known CFS-related genes (NR3C1 and COMT). For each combination of 20,160 genes and 41 SNPs, the best causal relationship was detected via causal model selection at Step 1. Numbers in parenthesis indicate the numbers of genes having each causal relationship with each SNP and disease status. At Step 2, each gene–SNP combination data was analyzed based on one of the three statistical models, corresponding to the detected causal relationship. Outside parenthesis, we present the numbers of significant genes identified by the corresponding statistical models. Three SNPs, each of which involves significant causal relationships with expression levels of more than 100 genes, are marked in bold.

[a] Logistic regression was conducted to identify genes whose expressions have interaction effect with genotype variation on disease status.
[b] Two-way ANOVA was conducted to identify genes whose expressions are affected by interaction between genotype variation and disease status.
[c] Independent test was conducted to identify genes whose expressions differ according to SNP genotypes.
[d] Glucocorticoid receptor located at 5q34.
[e] Catechol-O-methyltransferase located at 22q11.1.

Table 3
Significantly regulated pathways for SNP rs258750.

Pathway[a]	Model[b]	Source[c]	Nodes[d]	Gene ID[e]	Gene name
Galactose metabolism	Independent	KEGG	2/22	B4GALT2	UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 2
				MGAM	Maltase-glucoamylase
Basic mechanisms of SUMOylation	Independent	BioCarta	1/4	SUMO3	SMT3 suppressor of mif two 3 homolog 3
Internal ribosome entry pathway	Independent	BioCarta	1/8	EIF4E	Eukaryotic translation initiation factor 4E
Neutrophil and its surface molecules	Independent	BioCarta	1/8	ITGB2	Integrin, beta 2
Alternative complement pathway	Independent	BioCarta	1/9	CFB	Complement factor B
Mechanisms of protein import into the nucleus	Independent	BioCarta	1/9	NUP62	Nucleoporin 62kDa
Polyadenylation of mRNA	Independent	BioCarta	1/9	PABP2	Poly(A) binding protein II
B Lymphocyte cell surface molecules	Independent	BioCarta	1/9	ITGB2	Integrin, beta 2
Adhesion molecules on lymphocyte	Independent	BioCarta	1/9	ITGB2	Integrin, beta 2

Pathway enrichment analysis was conducted using 105 candidate independent genes, which were identified for rs258750. Significant biological pathways were detected via Fisher's exact test at a 5% significance level. Pathways are listed in order of significance, e.g., most significant pathway presents at the top.
[a] Name of biological pathway selected by Fisher's exact test.
[b] Causality models selected at Step 1.
[c] Source of pathway: Gene Map Annotator and Pathway Profiler (GenMapp), Kyoto Encyclopedia of Genes and Genomics (KEGG) and BioCarta.
[d] The number of candidate causal/reactive genes associated with pathway/the number of all genes associated with pathway.
[e] Gene ID of candidate causal/reactive genes associated with pathway.

immunity [19], because the gene expression data were obtained from a mononuclear cell, and the role of glucocorticoid receptor (NR3C1) gene is to regulate the level of glucocorticoid.

In the integrated analysis for comparing CFS-MDD/m with NF groups, for the rs6188 marker in the NR3C1 gene, 52 genes showed significant interaction effects with the rs6188 marker on disease status CFS-MDD/m from the logistic regression model. Also, the two-way ANOVA models yielded 217 candidate reactive genes, on which there are significant interaction effects between disease status and genotypes. Note that these candidate genes, especially reactive

Table 4
Significantly regulated pathways for SNP rs6188.

Pathway[a]	Model[b]	Source[c]	Nodes[d]	Gene ID[e]	Gene name
Electron transport chain	Causal	GenMapp	2/105	COX11	Cytochrome *c* oxidase subunit 11
				COX6A1	Cytochrome *c* oxidase subunit Via polypeptide 1
Steroid biosynthesis	Causal	GenMapp	1/9	F13B	Coagulation factor XIII, B polypeptide
Blood clotting cascade	Causal	GenMapp	1/19	F13B	Coagulation factor XIII, B polypeptide
FAS signaling pathway (CD95)	Causal	BioCarta	1/30	CFLAR	CASP8 and FADD-like apoptosis regulator
Induction of apoptosis through DR3 and DR4/5 Death Receptors	Causal	BioCarta	1/32	CFLAR	CASP8 and FADD-like apoptosis regulator
IL-2 receptor beta chain in T cell activation	Causal	BioCarta	1/35	CFLAR	CASP8 and FADD-like apoptosis regulator
HIV-1 Nef: negative effector of FAS and TNF	Causal	BioCarta	1/57	CFLAR	CASP8 and FADD-like apoptosis regulator
Agrin in postsynaptic differentiation	Reactive	BioCarta	3/39	UTRN	Utrophin
				DVL1	Dishevelled, dsh homolog 1
				ARHGEF6	Rac/Cdc42 guanine nucleotide exchange factor (GEF) 6
Cell cycle	Reactive	GenMapp	4/87	CDC14A	CDC14 cell division cycle 14 homolog A
				E2F2	E2F transcription factor 2
				CDC20	CDC20 cell division cycle 20homolog
Eicosanoid metabolism	Reactive	BioCarta	2/20	PTGES	Prostaglandin E synthase
				EPHX1	Epoxide hydrolase
Biosynthesis of cysteine	Reactive	BioCarta	1/2	CBS	Cystathionine-beta-synthase
Biosynthesis of threonine and methionine	Reactive	BioCarta	1/2	CBS	Cystathionine-beta-synthase
Inactivation of Gsk3 by AKT causes accumulation of β-catenin in alveolar macrophages	Reactive	BioCarta	2/25	MYD88	Myeloid differentiation primary response gene (88)
				DVL1	Dishevelled, dsh homolog 1
Fatty acid metabolism	Reactive	KEGG	3/57	HADHB	Hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase, beta subunit
				ADH6	Alcohol dehydrogenase 6 (class V)
				CYP19A1	Cytochrome *P450*, family 19, subfamily A, polypeptide 1
Bile acid biosynthesis	Reactive	KEGG	2/26	HADHB	Hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase, beta subunit
				ADH6	Alcohol dehydrogenase 6
Catabolic pathways for methionine, isoleucine, threonine and valine	Reactive	BioCarta	1/4	CBS	Cystathionine-beta-synthase
Basic mechanisms of SUMOylation	Reactive	BioCarta	1/4	SMT3H1	SMT3 suppressor of mif two 3 homolog 3
ALK in cardiac myocytes	Reactive	BioCarta	2/34	DVL1	Dishevelled, dsh homolog 1
				CHRD	Chordin
Taurine and hypotaurine metabolism[f]	Reactive	KEGG	1/5	GAD1	Glutamate decarboxylase 1
Biosynthesis of neurotransmitters[f]	Reactive	BioCarta	1/6	GAD1	Glutamate decarboxylase 1

Pathway enrichment analysis was conducted using 52 candidate causal genes and 217 candidate reactive genes, which were identified for rs6188. Significant biological pathways were detected via Fisher's exact test at a 5% significance level. Pathways are listed in order of significance within each of causality models, e.g., most significant pathway presents at the top.
[a] Name of biological pathway selected by Fisher's exact test.
[b] Causality models selected at Step 1.
[c] Source of pathway: Gene Map Annotator and Pathway Profiler (GenMapp), Kyoto Encyclopedia of Genes and Genomics (KEGG) and BioCarta.
[d] The number of candidate causal/reactive genes associated with pathway/the number of all genes associated with pathway.
[e] Gene ID of candidate causal/reactive genes associated with pathway.
[f] Pathways with *p*-value that is slightly larger than 0.05.

genes, could not be detected by Schadt et al.'s method. The proposed two-step integration method revealed the causal association among gene expression level, genotype and disease status in depth. Candidate causal/reactive genes were detected also for rs852977 in the NR3C1 gene. However, the candidate gene set for the rs852977 is very similar to that for the rs6188, with slight differences in causality structure. This similarity would be due to a strong linkage between the two SNPs.

Pathway enrichment analysis

In comparing CFS with NF groups, we conducted a pathway enrichment analysis of 105 genes that were identified to have a significant relationship with the rs258750 marker from the independent test at Step 2. The pathway classification showed that nine different pathways were associated with the rs258750 marker at the 5% significance level (Table 3). Out of nine pathways, four were enriched with genes involved in regulation of transcription, translation or mRNA processing, and three are related with immune system.

In order to compare CFS-MDD/m with NF groups, pathway enrichment analyses were conducted on the genes that were identified to have a significant relationship with the rs6188 and/or rs852977 markers at Step 2. Because of the linkage between the two SNPs, the results were similar (Tables 4 and 5), and we present herein details of the results only for the rs6188. While seven different pathways were detected at the 5% significance level for the 52 candidate causal genes, eleven different pathways were detected for the 217 candidate reactive genes (Table 4). In addition, two other pathways, whose *p*-values are slightly larger than the 5% significance level, are listed.

In pathway enrichment analyses of the candidate causal genes, the steroid biosynthesis pathway appears to have a direct causal effect on the disease status, CFS-MDD/m, through an integrative action of the rs6188 marker within the NR3C1 gene. The two significantly enriched biological pathways (i.e., 'IL-2 Receptor Beta Chain in T cell Activation', and 'HIV-1 Nef: negative effector of FAS and TNF') are all related to the immune system. On the other hand, the pathway enrichment analysis of the candidate reactive genes showed that several pathways related to lipid metabolism or biosynthesis, such as eicosanoid and fatty acid, appear to be important for responding to CFS-MDD/m. Furthermore, other pathways associated with neuron physiology and neurotransmitters appear to respond to CFS-MDD/m.

Discussion

We proposed a two-step procedure to integrate gene expression data, genotype variation data and clinical data, and to identify the genetic mechanism of a complex disease. We considered three different statistical tests based on the proposed two-step procedure for the combined data. For purposes of comparison, two different CFS-related datasets were analyzed via the multi-step procedure proposed by Schadt et al. [10] as well as based on our proposed method. In these specific datasets, no significant results were detected from the multi-step method of Schadt et al., while our method enabled us to identify many statistically significant causal relationships, some of which were biologically supported by pathway enrichment analyses. These results demonstrate that our method based on an exhaustive search investigation would have more power, and motivates us to study the power of our proposed method via simulations in the future.

Furthermore, our proposed approach provided some interesting results. First, CFS groups and CFS-MDD/m groups would appear to have different genotypes and gene expression profiles even though they had the common characteristic of chronic fatigue. In particular, CFS has major susceptibility markers within the NR3C1 gene, and CFS-MDD/m seems to have major susceptibility markers within the catechol-O-methyltransferase (COMT) gene, though they are not statistically significant after FDR correction (Table 1B). The NR3C1 gene regulates the level of glucocorticoid which is the end product of the hypothalamic-pituitary-adrenal (HPA) whereas COMT catalyzes the transfer of a methyl group from S-adenosylmethionine to catecholamines, which is the principal end product of the sympathetic nervous system (SNS), whose role is maintaining stress-related homeostasis [20]. The different major susceptibility gene may be related with to the provoking of MDD/m.

Second, polymorphisms in the glucocorticoid receptor NR3C1 gene act on CFS and CFS-MDD/m differently. The polymorphisms (rs258750) within NR3C1 have significant effects on CFS, and the 105 gene expression levels independently. However, in the integrated analysis for comparing CFS-MDD/m and NF groups, polymorphisms within the NR3C1 gene affect the CFS-MDD/m and several gene

Table 5
Significantly regulated pathways for SNP rs852977.

Pathway[a]	Model[b]	Source[c]	Nodes[d]	Gene ID[e]	Gene name
Agrin in postsynaptic differentiation	Casual	BioCarta	2/39	DMD	Dystrophin
				DVL1	Dishevelled, dsh homolog 1
Steroid biosynthesis	Casual	GenMAPP	1/9	F13B	Coagulation factor XIII, B polypeptide
Nucleotide GPCRs	Casual	GenMAPP	1/10	P2RY4	Pyrimidinergic receptor P2Y, G-protein coupled 4
RNA polymerase III transcription	Casual	BioCarta	1/8	GTF3C1	General transcription factor IIIC, polypeptide 1, alpha 220kDa
Blood clotting cascade	Casual	GenMAPP	1/19	F13B	Coagulation factor XIII, B polypeptide
Bile acid biosynthesis	Casual	KEGG	1/26	ADH6	Alcohol dehydrogenase 6
Tyrosine metabolism	Casual	KEGG	1/37	ADH6	Alcohol dehydrogenase 6
Inactivation of Gsk3 by AKT causes accumulation of b-catenin in alveolar macrophages	Reactive	BioCarta	2/25	MYD88	Myeloid differentiation primary response gene (88)
				DVL1	Dishevelled, dsh homolog 1
ALK in cardiac myocytes	Reactive	BioCarta	2/34	DVL1	Dishevelled, dsh homolog 1
				CHRD	Chordin
Biosynthesis of neurotransmitter	Reactive	BioCarta	1/6	GAD1	Glutamate decarboxylase 1
Taurine and hypotaurine metabolism	Reactive	KEGG	1/5	GAD1	Glutamate decarboxylase 1
Electron transport chain	Reactive	GenMAPP	2/105	COX11	Cytochrome *c* oxidase subunit 11
				COX6A1	Cytochrome *c* oxidase subunit VIa polypeptide 1

Pathway enrichment analysis was conducted using 120 candidate causal genes and 73 candidate reactive genes, which were identified for rs852977. Significant biological pathways were detected via Fisher's exact test at a 5% significance level. Pathways are listed in order of significance within each of causality models, e.g., most significant pathway presents at the top.

[a] Name of biological pathway selected by Fisher's exact test.
[b] Causality models selected at Step 1.
[c] Source of pathway: Gene Map Annotator and Pathway Profiler (GenMapp), Kyoto Encyclopedia of Genes and Genomics (KEGG) and BioCarta.
[d] The number of candidate causal/reactive genes associated with pathway/the number of all genes associated with pathway.
[e] Gene ID of candidate causal/reactive genes associated with pathway.

expression levels differently. For example, the 217 genes are differentially expressed according to the rs6188 marker genotype within NR3C1 and disease status, even though polymorphisms within NR3C1 have no direct significant effects after FDR correction on the CFS-MDD/m. In addition, the 52 genes also regulate the CFS-MDD/m, through integrated action with the rs6188 marker. The different action of the NR3C1 gene on gene expression level and disease may be an outcome of other factors, such as environmental effects or polymorphisms of the COMT gene. The catecholamines which are regulated by the COMT gene, have been often been regarded as immunosuppressive [20].

Two pathway enrichment analyses for the 52 candidate causal genes and 217 candidate reactive genes indicated that our approach can recover plausible regulatory mechanisms of CFS-MDD/m by comparing CFS-MDD/m and NF groups. From the comparison, we noticed that the pathways related to the immune system and steroid may have causal effect on disease state through an integrative action of the NR3C1 gene. Both the NR3C1 gene that regulates the level of glucocorticoid, and the steroid that includes corticosteroids are known to regulate the immune function [19]. A number of studies have found many irregularities in the immune systems in CFS patients [21]. This suggested that an important cause of CFS-MDD/m would be the immune system dysfunction, regulated by the neuroendocrine system, which rs6188 in the NR3C1 gene seems influence. Another potential implication of this comparison is that the CFS-MDD/m status and genetic polymorphisms can jointly induce different activation and expression of several lipid related metabolisms, neuron physiology differentiation, and neurotransmitters. Our results can be supported by the known relationship between eicosanoid or fatty acid and CFS [22–25].

However, since fatigue is a core symptom in major depressive disorder [26], CFS-MDD/m patients might have fatigue due to the depression rather than unexplained causes, and hence the significant results may be related to a 'major depression disorder with melancholic features' rather than chronic fatigue. For example, the excessive hypothalamus-pituitary-adrenal (HPA) axis responses, of which end products are glucocorticoids, are known to be hallmarks of depression [27–29]. Also, the major depression can be associated with the immune activation, dysfunction of neurotransmitters at synapse [30–32], and essential fatty acids [33].

Our integrative analyses considering the interaction effect among different levels of data could elucidate different disease susceptibility and differentially expressed genes of genetically different individuals. Some results showed that integrating genotype and expression data may help the search for new directions for the treatment of common human diseases that are not being detected using only one type of data. The integrated analysis provided more information than the two separate analyses of gene expression data and genotype variation data for characterizing CFS that has several possible causes.

In conclusion, our approach to the integration of heterogeneous data sets can be generally applied to other studies in which gene expression data, genotype variation data and clinical data are available, and it can be very useful as the importance of integrated data analysis has been increasing. The proposed approach can also be extended to datasets containing other type of data, such as protein data rather than clinical data. Our approach can be readily applicable to quantitative traits rather than binary clinical outcome traits, by employing linear regression analysis. Also, it can be easily applied to genome-wide association studies, and can handle environmental factors, such as age and sex, by treating these factors as covariates in the regression model. Furthermore, our approach can be extended to the gene-set approach, the module based approach or co-expression network as Presson et al. [14] and Chen et al. [16] did.

However, there are some limitations to the proposed method. First, the causality models, which we assumed herein, are too simple to represent true mechanisms, which would be more complicated due to possible interactions between causal-reactive genes [10]. Further considerations for more complicated models are necessary in order to identify the genetic mechanism of complex diseases. Second, the proposed approach may need large computing although it is applicable to genome-wide studies because it is not limited in the scale of data. Another limitation would be a misclassification problem in that the proposed method relies on the LCMS. Our current approach does not use FDR procedure to account for the model misclassification problem. In fact, FDR procedure was employed only in the second step, not in the first step for the model selection procedure that chooses the model with the minimum AIC among the three causal models. While anticipating the problem, we still employed the LCMS process because it showed good power for detecting true models in the simulation evaluated by Schadt et al. [10]. Our approach can be extended to account for the errors caused by the model misclassification in the first step. For example, we can test for the difference in the AIC values of three causality models, because the chance for model misclassification would be high when the difference between the smallest AIC value from the selected model and those from the other models is not large. A permutation-based nonparametric test might be developed for this testing. We think it requires a further study to control simultaneously two types of errors: causality model selection, and significant maker-gene pair identification. Finally, we plan to address the multiplicity problem for tests with (tightly) linked markers, although we used a FDR procedure for lack of standard methods that deal with multiplicity among dependent tests.

Methods

Chronic fatigue syndrome (CFS) dataset

Chronic fatigue syndrome (CFS) is a debilitating illness lacking consistent anatomic lesions and eluding conventional laboratory diagnosis. CFS has no confirmatory physical signs or laboratory abnormalities, and its etiology and pathophysiology are unknown. This disease characterized by severe chronic fatigue, lasting at least 6 months, which is accompanied by symptoms such as impairment in short-term memory or concentration, sore throat, tender lymph nodes, and muscle pain [34]. The CDC Chronic Fatigue Syndrome Research Group produced the dataset for 173 patients including gene expression, proteomic, single nucleotide polymorphism (SNP), and clinical data, and the dataset is available to the public online (*http://www.camda.duke.edu/camda06/datasets/index.html*).

The CAMDA 2006 competition datasets were used for the analysis in this study. The patients include 34 males and 137 females, and the majority race is white. According to severity of symptoms, the patients were classified into five groups of CFS. In this study, we only consider three groups: 46 patients meeting the CFS research case definition (CFS), 19 patients meeting the CFS research case definition and having 'a major depressive disorder with melancholic features' (CFS-MDD/m), and 36 patients who show no fatigue (NF).

This CFS data set has been analyzed by many research groups for identifying molecular markers and elucidating pathophysiology of CFS [35], for finding two differentially expressed genes related with fatigue and depression, respectively [36–38], for discriminating classes of unexplained chronic fatigue based on differential gene expressions [39,40], and for examining the relationship between CFS and allostatic load based on the clinical dataset [41–46]. In the CFS dataset, the expression levels of 20,160 genes were assessed from peripheral blood mononuclear cells, via custom-printed single-channel oligonucleotide chips. We conducted quantile normalization [47] on the gene expression data which were pre-processed by the original CDC research group. For genotype data, the whole blood DNA was extracted and specific areas of the genes of interest were amplified by PCR [48–51].

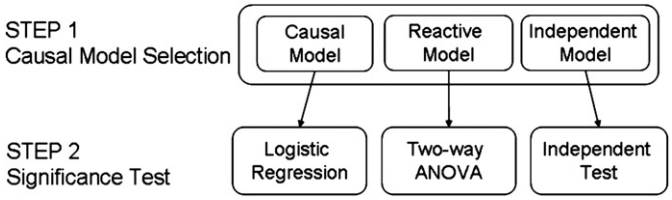

Fig. 1. Flow diagram to illustrate two-step integration procedure based on causal relationship. In the first step, for each gene expression-genetic variation combination, the most appropriate causality models are selected to understand the direction of the causal relationship among genetic variation, gene expression level, and disease. In the second step, significance testing is carried out based on a statistical model for each combination, such as logistic regression and a two-way analysis of variance (ANOVA), according to the causality model selected from the first step.

Two-step procedure for integrative analysis

A two-step procedure was developed to integrate gene expression, genotype data and clinical data, and thus to elucidate mechanisms underlying disease susceptibility and progression. In Fig. 1, a flow diagram is presented to illustrate the integration method based on causal relationship among the three different levels of data. In the first step, the most appropriate causality models are selected to understand the direction of the causal relationship among genetic variation, gene expression level, and disease for each gene expression-genetic variation combination. In the second step, significance testing is carried out based on a statistical model for each combination, such as logistic regression and a two-way analysis of variance (ANOVA), according to the causality model selected from the first step. Via these tests, we identify gene expression traits whose effects on disease status or responses to disease status are modified by the genotype variation effects.

Step 1: Causality model selection

The possible causal relationships among genetic variation, gene expression level and disease trait can be represented by graphical models [10]. Three simple models, such as causal, reactive and independent models, are illustrated in Fig. 2. In this study, we assume that each pair of genetic locus and expressed gene has one of these three simple causal relationships, according to which we build statistical models to examine potential relationships among the gene expression level, genotype variation and disease status. In order to find the most possible causal relationship among the three heterogeneous data, we adapted the likelihood based causality model selection (LCMS) test [10], which uses conditional correlation measures for determining the relationships best supported by the data. First, likelihoods associated with each of the causality models are constructed and maximized with respect to the model parameters. Second, the best model is selected for each SNP-transcript combination, by using Akaike Information Criterion (AIC [52]) which can be used to compare different models.

Assuming standard Markov properties for the simple graphs, the joint probability distributions for the three models are as follows:

- Causal Model: $P(S, R, D) = P(S)\,P(R|S)\,P(D|R)$
- Reactive Model: $P(S, R, D) = P(S)\,P(D|S)\,P(R|D)$
- Independent Model: $P(S, R, D) = P(S)\,P(R|S)\,P(D|R, S)$

Here S represents a genotype variation, R gene expression, and D disease status. $P(S)$ is the genotype probability distribution for marker S and is further assumed to be co-dominant. $P(R|S)$ and $P(R|D)$ are the conditional probabilities of R given genotypes S and disease status D, respectively. In an application to the CFS data, we further assume that the random variable R is conditionally normally distributed, and the random variable D has a binomial distribution. Therefore, in probability $P(D|R)$, the random variable D has a binominal distribution with a success probability that can be modelled by a logistic regression model. $P(D|S)$ is the probability distribution of D conditional on locus S, in which the random variable D also has a binomial distribution. Based on these assumptions, the likelihood of a correspondence to each of the joint probability distributions is constructed. For each model, the model parameters are estimated via a standard maximum likelihood method. The best model supported by the data is chosen based on the AIC, which is commonly used to compare models with different numbers of parameters.

Step 2: Statistical tests for identifying candidate causal and reactive genes

In Step 2, we perform statistical tests to determine the significance of the genetic regulatory relationships described in the causality model that are selected at Step 1. The determination of response and independent variables in the statistical models depend on the causality model selected at Step 1. These statistical tests allow us to consider the interaction effects among the three different levels of data and to elucidate differences in disease susceptibility and gene expression pattern across genetically different individuals. The two-step procedure results in a set of candidate causal and reactive genes, whose expressions affect disease status and respond to disease status under the influence of genotype variation, respectively.

Test for causal model

In order to investigate gene expression traits whose effects on disease status are modified by genotype variation, we examine the interaction effect of genotype variation and gene expression level on the disease status using logistic regression below:

$$\text{log it}(\pi) = S + R + S \times R, \tag{1}$$

where π represents the probability of getting the disease; S represents the effect of genotype variation such as SNPs; R represents the effect of gene expression levels; and $S \times R$ represents the interaction effect between genotype variation and gene expression level.

Test for reactive model

For investigating gene expression traits whose responses to disease status are affected by genotype variations, we fit the following two-way ANOVA model with the interaction between genotype variation and disease groups:

$$R = S + D + S \times D, \tag{2}$$

where S represents the effect of genotype variation; D represents the effect of disease groups; and $S \times D$ represents the interaction effect between genotype variation and disease groups.

Test for independent model

When the independent model is selected at Step 1, the effect of genotype variation on each of gene expression and disease can be

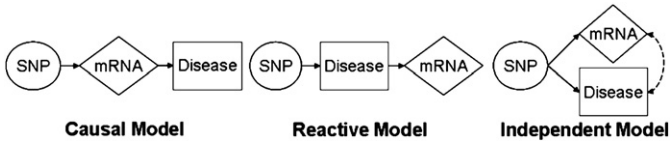

Fig. 2. Possible causal relationships among mRNA level, genotype variation and complex disease proposed by Schadt et al. [10]. SNP represents genotype variation, mRNA indicates the expression level of a gene, and disease represents a phenotype of interest, such as a complex disease. Three possible simplest causal relationships are represented, given that mRNA and disease are somehow under the control of a common SNP.

investigated separately. First, the logistic regression is employed to detect genotypic markers linked to disease loci:

$$\log it(\pi) = S \quad (3)$$

Next, we identify genotypic markers that regulate gene expression levels, based on the one-way ANOVA model where the dependent variable is R and the independent variable is S.

Detecting significance of association via multiple testing adjustment

In the Step 2, significant associations among genotype variation, gene expression and disease are declared via statistical tests for all possible pairs of gene expression-genotype variation. Due to the large number of tests, the multiple testing problem needs to be addressed. In order to adjust this multiplicity, we used a step-up procedure controlling false discovery rate (FDR [18]) of 5%.

Pathway enrichment analysis

Pathway enrichment analysis was performed on each of the candidate causal/reactive gene sets selected from two-step procedure to identify 'significantly regulated pathways' which are defined as significantly over-presented pathways for a particular gene set. We used the pathway annotations in Gene Map Annotator and Pathway Profiler (GenMapp) [53], Kyoto Encyclopaedia of Genes and Genomics (KEGG) [54], and BioCarta (*http://www.biocarta.com/*) when employing software called ArrayXPath (*http://www.snubi.org/software/ArrayXPath/*) [55] for the pathway enrichment analysis. This study involves comparing 'the proportion of genes in a specific pathway' among a list of differentially regulated genes detected from our two-step integration to that among all genes in a database. For the comparison, the association between a particular pathway and a specific list of genes is tested by constructing a 2×2 contingency table. For the statistical significance of pathway enrichments, p-values were calculated by the Fisher's exact test [56] based on a hypergeometric distribution. Note that this approach has been widely used in functional interpretation of gene expression data analysis using gene ontology (GO) enrichment analysis [57,58] and pathway enrichment analysis [59,60].

Acknowledgments

The authors thank three anonymous reviewers for their valuable comments. The work was supported by the National Research Laboratory Program of Korea Science and Engineering Foundation (M10500000126) and the Brain Korea 21 Project of the Ministry of Education.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2009.06.002.

References

[1] T.R. Hughes, M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H.Y. Dai, Y.D.D. He, M.J. Kidd, A.M. King, M.R. Meyer, D. Slade, P.Y. Lum, S.B. Stepaniants, D.D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, S.H. Friend, Functional discovery via a compendium of expression profiles, Cell 102 (2000) 109–126.
[2] C.L. Karp, A. Grupe, E. Schadt, S.L. Ewart, M. Keane-Moore, P.J. Cuomo, J. Kohl, L. Wahl, D. Kuperman, S. Germer, D. Aud, G. Peltz, M. Wills-Karp, Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma, Nat. Immunol. 1 (2000) 221–226.
[3] K. Ozaki, R. Spolski, C.G. Feng, C.F. Qi, J. Cheng, A. Sher, H.C. Morse, C.Y. Liu, P.L. Schwartzberg, W.J. Leonard, A critical role for IL-21 in regulating immunoglobulin production, Science 298 (2002) 1630–1634.
[4] R.J. Klein, C. Zeiss, E.Y. Chew, J.Y. Tsai, R.S. Sackler, C. Haynes, A.K. Henning, J.P. SanGiovanni, S.M. Mane, S.T. Mayne, M.B. Bracken, F.L. Ferris, J. Ott, C. Barnstable, J. Hoh, Complement factor H polymorphism in age-related macular degeneration, Science 308 (2005) 385–389.
[5] G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 5116–5121.
[6] P. Baldi, A.D. Long, A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes, Bioinformatics 17 (2001) 509–519.
[7] B. Efron, R. Tibshirani, Empirical Bayes methods and false discovery rates for microarrays, Genet. Epidemiol. 23 (2002) 70–86.
[8] D.W. Hosmer, S. Lemeshow, E.D. Cook, Applied Logistic Regression, Wiley, New York ; Chichester, 2000.
[9] E.J. Chesler, L. Lu, S.M. Shou, Y.H. Qu, J. Gu, J.T. Wang, H.C. Hsu, J.D. Mountz, N.E. Baldwin, M.A. Langston, D.W. Threadgill, K.F. Manly, R.W. Williams, Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function, Nat. Genet. 37 (2005) 233–242.
[10] E.E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S.K. Sieberts, S. Monks, M. Reitman, C.S. Zhang, P.Y. Lum, A. Leonardson, R. Thieringer, J.M. Metzger, L.M. Yang, J. Castle, H.Y. Zhu, S.F. Kash, T.A. Drake, A. Sachs, A.J. Lusis, An integrative genomics approach to infer causal associations between gene expression and disease, Nat. Genet. 37 (2005) 710–717.
[11] S.A. Monks, A. Leonardson, H. Zhu, P. Cundiff, P. Pietrusiak, S. Edwards, J.W. Phillips, A. Sachs, E.E. Schadt, Genetic inheritance of gene expression in human cell lines, Am. J. Hum. Genet. 75 (2004) 1094–1105.
[12] M. Morley, C.M. Molony, T.M. Weber, J.L. Devlin, K.G. Ewens, R.S. Spielman, G. Cheung, Genetic analysis of genome-wide variation in human gene expression, Nature 430 (2004) 743–747.
[13] C.M. Kendziorski, M. Chen, M. Yuan, H. Lan, A.D. Attie, Statistical methods for expression quantitative trait loci (eQTL) mapping, Biometrics 62 (2006) 19–27.
[14] A.P. Presson, E.M. Sobel, J.C. Papp, C.J. Suarez, T. Whistler, M.S. Rajeevan, S.D. Vernon, S. Horvath, Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome, BMC Syst. Biol. 2 (2008) 95.
[15] B. Zhang, S. Horvath, A general framework for weighted gene co-expression network analysis, Stat. Appl. Genet. Mol. Biol. 4 (2005) Article17.
[16] Y. Chen, J. Zhu, P.Y. Lum, X. Yang, S. Pinto, D.J. MacNeil, C. Zhang, J. Lamb, S. Edwards, S.K. Sieberts, A. Leonardson, L.W. Castellini, S. Wang, M.F. Champy, B. Zhang, V. Emilsson, S. Doss, A. Ghazalpour, S. Horvath, T.A. Drake, A.J. Lusis, E.E. Schadt, Variations in DNA elucidate molecular networks that cause disease, Nature 452 (2008) 429–435.
[17] H. Lan, M. Chen, J.B. Flowers, B.S. Yandell, D.S. Stapleton, C.M. Mata, E.T. Mui, M.T. Flowers, K.L. Schueler, K.F. Manly, R.W. Williams, C. Kendziorski, A.D. Attie, Combined expression trait correlations and expression quantitative trait locus mapping, PLoS Genet. 2 (2006) e6.
[18] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate — a practical and powerful approach to multiple testing, J. R. Stat. Soc., B 57 (1995) 289–300.
[19] J.I. Webster, L. Tonelli, E.M. Sternberg, Neuroendocrine regulation of immunity, Annu. Rev. Immunol. 20 (2002) 125–163.
[20] J. Elenkov, R.L. Wilder, G.P. Chrousos, E.S. Vizi, The sympathetic nerve — an integrative interface between two supersystems: the brain and the immune system, Pharmacol. Rev. 52 (2000) 595–638.
[21] B.H. Natelson, M.H. Haghighi, N.M. Ponzio, Evidence for the presence of immune dysfunction in chronic fatigue syndrome, Clin. Diagn. Lab. Immunol. 9 (2002) 747–752.
[22] J.B. Gray, A.M. Martinovic, Eicosanoids and essential fatty-acid modulation in chronic disease and the chronic fatigue syndrome, Med. Hypotheses 43 (1994) 31–42.
[23] B.K. Puri, Long-chain polyunsaturated fatty acids and the pathophysiology of myalgic encephalomyelitis (chronic fatigue syndrome), J. Clin. Pathol. 60 (2007) 122–124.
[24] B.K. Puri, J. Holmes, G. Hamilton, Eicosapentaenoic acid-rich essential fatty acid supplementation in chronic fatigue syndrome associated with symptom remission and structural brain changes, Int. J. Clin. Pract. 58 (2004) 297–299.
[25] Z. Liu, D. Wang, Q. Xue, J. Chen, Y. Li, X. Bai, L. Chang, Determination of fatty acid levels in erythrocyte membranes of patients with chronic fatigue syndrome, Nutr. Neurosci. 6 (2003) 389–392.
[26] C.U. Pae, H.K. Lim, C. Han, A.A. Patkar, D.C. Steffens, P.S. Masand, C. Lee, Fatigue as a core symptom in major depressive disorder: overview and the role of bupropion, Expert Rev. Neurotherapeutics 7 (2007) 1251–1263.
[27] C.M. Pariante, A.H. Miller, Glucocorticoid receptors in major depression: relevance to pathophysiology and treatment, Biol. Psychiatry 49 (2001) 391–404.
[28] F. Holsboer, The corticosteroid receptor hypothesis of depression, Neuropsychopharmacology 23 (2000) 477–501.
[29] C.M. Pariante, Glucocorticoid receptor function in vitro in patients with major depression, Stress 7 (2004) 209–219.
[30] A. Neumeister, T. Young, J. Stastny, Implications of genetic research on the role of the serotonin in depression: emphasis on the serotonin type 1(A) receptor and the serotonin transporter, Psychopharmacology 174 (2004) 512–524.
[31] G. Sanacora, R. Gueorguieva, C.N. Epperson, Y.T. Wu, M. Appel, D.L. Rothman, J.H. Krystal, G.F. Mason, Subtype-specific alterations of gamma-aminobutyric acid and glutamate in patients with major depression, Arch. Gen. Psychiatry 61 (2004) 705–713.
[32] M. Maes, H.Y. Meltzer, The Serotonin Hypothesis of Major Depression, Raven Press, New York, 1995.
[33] A.C.P. Van Strater, P.F. Bouvy, Omega-3 fatty acids and mood disorders, Am. J. Psychiatr. 163 (2006) 2018.
[34] K. Fukuda, S.E. Straus, I. Hickie, M.C. Sharpe, J.G. Dobbins, A. Komaroff, A. Schluederberg, J.F. Jones, A.R. Lloyd, S. Wessely, N.M. Gantz, G.P. Holmes, D. Buchwald, S. Abbey, J. Rest, J.A. Levy, H. Jolson, D.L. Peterson, J.H.M.M. Vercoulen, U. Tirelli, B. Evengard, B.H. Natelson, L. Steele, M. Reyes, W.C. Reeves, The chronic

fatigue syndrome — a comprehensive approach to its definition and study, Ann. Intern. Med. 121 (1994) 953–959.
[35] S.D. Vernon, W.C. Reeves, The challenge of integrating disparate high-content data: epidemiological, clinical, and laboratory data collected during an in-hospital study of chronic fatigue syndrome, Pharmacogenomics 7 (2006) 345–354.
[36] H. Fang, Q. Xie, R. Boneva, J. Fostel, R. Perkins, W.D. Tong, Gene expression profile exploration of a large dataset on chronic fatigue syndrome, Pharmacogenomics 7 (2006) 429–440.
[37] G. Broderick, R.C. Craddock, T. Whistler, R. Taylor, N. Klimas, E.R. Unger, Identifying illness parameters in fatiguing syndromes using classical projection methods, Pharmacogenomics 7 (2006) 407–419.
[38] J. Fostel, R. Boneva, A. Lloyd, Exploration of the gene expression correlates of chronic unexplained fatigue using factor analysis, Pharmacogenomics 7 (2006) 441–454.
[39] L. Carmel, S. Efroni, P.D. White, E. Aslakson, U. Vollmer-Conna, M.S. Rajeevan, Gene expression profile of empirically delineated classes of unexplained chronic fatigue, Pharmacogenomics 7 (2006) 375–386.
[40] T. Whistler, R. Taylor, R.C. Craddock, G. Broderick, N. Klimas, E.R. Unger, Gene expression correlates of unexplained fatigue, Pharmacogenomics 7 (2006) 395–405.
[41] E.M. Maloney, B.M. Gurbaxani, J.F. Jones, L.D. Coelho, C. Pennachin, B.N. Goertzel, Chronic fatigue syndrome and high allostatic load, Pharmacogenomics 7 (2006) 467–473.
[42] B.N. Goertzel, C. Pennachin, L.D. Coelho, E.M. Maloney, J.F. Jones, B. Gurbaxani, Allostatic load is associated with symptoms in chronic fatigue syndrome patients, Pharmacogenomics 7 (2006) 485–494.
[43] B.M. Gurbaxani, J.F. Jones, B.N. Goertzel, E.M. Maloney, Linear data mining the Wichita clinical matrix suggests sleep and allostatic load involvement in chronic fatigue syndrome, Pharmacogenomics 7 (2006) 455–465.
[44] U. Vollmer-Conna, E. Aslakson, P.D. White, An empirical delineation of the heterogeneity of chronic unexplained fatigue in women, Pharmacogenomics 7 (2006) 355–364.
[45] E. Aslakson, U. Vollmer-Conna, P.D. White, The validity of an empirical delineation of heterogeneity in chronic unexplained fatigue, Pharmacogenomics 7 (2006) 365–373.
[46] R.C. Craddock, R. Taylor, G. Broderick, T. Whistler, N. Klimas, E.R. Unger, Exploration of statistical dependence between illness parameters using the entropy correlation coefficient, Pharmacogenomics 7 (2006) 421–428.
[47] B.M. Bolstad, R.A. Irizarry, M. Astrand, T.P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics 19 (2003) 185–193.
[48] J.M. Fernandez-Real, J. Vendrell, W. Ricart, M. Broch, C. Gutierrez, R. Casamitjana, J. Oriola, C. Richart, Polymorphism of the tumor necrosis factor-alpha receptor 2 gene is associated with obesity, leptin levels, and insulin resistance in young subjects and diet-treated type 2 diabetic patients, Diabetes Care 23 (2000) 831–837.
[49] J.M. Fernandez-Real, M. Broch, J. Vendrell, C. Gutierrez, R. Casamitjana, M. Pugeat, C. Richart, W. Ricart, Interleukin 6 gene polymorphism and insulin sensitivity, Diabetes 49 (2000) A396.
[50] A. Emptoz-Bonneton, P. Cousin, K. Seguchi, G.V. Avvakumov, C. Bully, G.L. Hammond, M. Pugeat, Novel human corticosteroid-binding globulin variant with low cortisol-binding affinity, J. Clin. Endocrinol. Metab. 85 (2000) 361–367.
[51] M.J. Arranz, J. Munro, J. Birkett, A. Bolonna, D. Mancama, M. Sodhi, K.P. Lesch, J.F.W. Meyer, P. Sham, D.A. Collier, R.M. Murray, R.W. Kerwin, Pharmacogenetic prediction of clozapine response, Lancet 355 (2000) 1615–1616.
[52] H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (1974) 716–723.
[53] K.D. Dahlquist, N. Salomonis, K. Vranizan, S.C. Lawlor, B.R. Conklin, GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways, Nat. Genet. 31 (2002) 19–20.
[54] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, The KEGG resource for deciphering the genome, Nucleic Acids Res. 32 (2004) D277–D280.
[55] H.J. Chung, M. Kim, C.H. Park, J. Kim, J.H. Kim, ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics, Nucleic Acids Res. 32 (2004) W460–W464.
[56] R.A. Fisher, The logic of inductive inference, J. R. Stat. Soc. 98 (1935) 39–54.
[57] B.R. Zeeberg, W.M. Feng, G. Wang, M.D. Wang, A.T. Fojo, M. Sunshine, S. Narasimhan, D.W. Kane, W.C. Reinhold, S. Lababidi, K.J. Bussey, J. Riss, J.C. Barrett, J.N. Weinstein, GoMiner: a resource for biological interpretation of genomic and proteomic data, Genome Biology 4 (2003) R28.
[58] T. Beissbarth, T.P. Speed, GOstat: find statistically overrepresented Gene Ontologies within a group of genes, Bioinformatics 20 (2004) 1464–1465.
[59] P. Grosu, J.P. Townsend, D.L. Hartl, D. Cavalieri, Pathway processor: a tool for integrating whole-genome expression results into metabolic networks, Genome Res. 12 (2002) 1121–1126.
[60] D.Y. Pan, N. Sun, K.H. Cheung, Z. Guan, L.G. Ma, M. Holford, X. Deng, H. Zhao, PathMAPA: a tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for *Arabidopsis*, BMC Bioinformatics 4 (2003) 56.